DOCUMENT RESUME

ED 457 174 TM 033 264

AUTHOR Haueisen, Heidi L.

TITLE Using Multiple Raters on Performance Based Driving Tests with High School Driver Education Students.

PUB DATE 2001-05-00

NOTE 46p.; Master of Arts Action Research Project, St. Xavier University and Skylight Professional Development.

PUB TYPE Dissertations/Theses (040)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *Driver Education; High School Students; High Schools; *Interrater Reliability; *Performance Based Assessment; *Student Evaluation; Test Construction

ABSTRACT

An assessment tool was designed and implemented to increase consistent application among and between multiple raters assessing students in driver education. The targeted population was students in grades 9 through 12 enrolled in drive education at a high school in an affluent suburb near a large city. The problem of a lack of a consistent assessment tool within the department was documented by anecdotal records of department meetings, surveys of teachers, and individual interviews with teachers. Analysis of the probable causes of the problem indicated the current lack of assessment instruments, and a lack of training or familiarity with the method as a major source of the inconsistency. The inconsistency affected interrater reliability and the consistency for the individual raters themselves. Review of the research indicated that raters easily introduce error into scores because of unfamiliarity or inadequate training related to the rating scale. A review of the solution strategies adopted by other researchers resulted in the development of a teacher-generated progress report for driver's education laboratory students to be updated three or more times per semester, a scoring rubric for lab students, and departmental workshops on proper implementation and consistent usage. Two driver's education classes of seven and eight students each were divided into four lab groups of two students, with one group consisting of only one student. Assessment sessions occurred every fourth day of the month-long summer session, with two raters independently observing and assessing the same lab session using the assessment instrument. The instrument showed high levels of consistent scoring between raters, and teachers and students appeared to benefit. (Contains 26 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

TM

ED 457 174

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

H. Haueisen

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

USING MULTIPLE RATERS ON PERFORMANCE BASED DRIVING TESTS WITH HIGH SCHOOL DRIVER EDUCATION STUDENTS

Heidi L. Haueisen

An Action Research Project Submitted to the Graduate Faculty of the

School of Education in Partial Fulfillment of the

Requirements for the Degree of Master of Arts in Teaching and Leadership

Saint Xavier University & Skylight Professional Development

Field-Based Master's Program

Chicago, Illinois

May, 2001

TM033264

BEST COPY AVAILABLE

SIGNATURE PAGE

This project was approved by

Dr. Susan L. Mason
Advisor

Brenda Barbknecht
Advisor

Shirley Colley
Dean, School of Education

DEDICATION

To BNH and BNP for keeping me dry throughout the storm
while maintaining my sense of humor.

In loving memory of my parents, Dar and Marty,
who continue to teach me life's greatest lessons,
neither seen nor heard,
but deeply felt.

Special thanks to Mike Regnier for his willingness to offer his
classes as research subjects and his contribution
to the data collection phase.
His participation and easy going nature throughout the
summer session resulted in an ideal atmosphere
in which to work.

Sincere appreciation to Rick Nelson for his technical support
and patience in assisting with the final phase of
the project layout. He was instrumental in
enhancing the appearance
of the paper.

ABSTRACT

This paper describes a program for designing and implementing an assessment tool to increase the consistent application among and between multiple raters. The targeted population consists of students in grades 9-12 and enrolled in Driver Education. The high school is located in an affluent suburban area, not far from a large metropolitan city. The problem of a lack of consistent assessment tool within the department was documented by anecdotal records of department meetings, surveys of teachers, as well as individual interviews with teachers.

Analysis of probable causes indicated the current lack of assessment instruments, as well as a lack of training or familiarity with the method as a major source of inconsistency for most teacher. This inconsistency affected inter-rater reliability as well as consistency among individual raters themselves. Review of the research indicated raters easily introduced error into scores due to unfamiliarity or inadequate training towards the rating scale.

A review of the solution strategies by other researchers resulted in a teacher generated progress report for lab students to be updated three or more times per semester, a scoring rubric for lab students and departmental workshops on proper implementation and consistent usage.

TABLE OF CONTENTS

CHAPTER 1 - PROBLEM STATEMENT AND CONTEXT 1

General Statement of the Problem 1

Immediate Problem Context 1

The Surrounding Community 3

National Context of the Problem 5

CHAPTER 2 - PROBLEM DOCUMENTATION 7

Problem Evidence 7

Probable Causes 8

CHAPTER 3 - THE SOLUTION STRATEGY 12

Literature Review 12

Project Objectives and Processes 22

Project Action Plan 22

Methods of Assessment 23

CHAPTER 4 - PROJECT RESULTS 24

Historical Description of the Intervention 24

Presentation and Analysis of Results 25

Conclusions and Recommendations 27

REFERENCES 30

APPENDICES

A DRIVER EDUCATION STAFF SURVEY 33

B REVISED ASSESSMENT INSTRUMENT 34

C (P. M. I.) SURVEY SHEET 35

D SAMPLE LETTER TO PARENTS 36

CHAPTER 1

PROBLEM STATEMENT AND CONTEXT

General Statement of the Problem

The teachers of the targeted driver education classes lack a consistent assessment instrument for the performance based behind-the-wheel phase of the student driver education program. Evidence for the existence of the problem includes anecdotal records of department meetings, interviews with teachers, and application of current assessment tools.

Immediate Problem Context

Demographic information pertaining to the high school site was gathered from the 1999 School Report Card. The school site researched is a high school, grades nine through twelve, which is by itself a single high school district. Total school enrollment in 1998 was 3,274 students, which places the site in the large high school category on the School Report Card results. The enrollment was reported as 85.8% White, 0.9% Black, 2.4% Hispanic, 10.9% Asian/Pacific Islander, and 0.0% Native American. One percent of the students were considered low-income, 1.6% are Limited-English proficient, and 0.2% are dropouts. Attendance rates reached 95.3% in 1998, with chronic truancy at 1.6%, and mobility at 6.3%. The number of chronic truants in the district was fifty-three.

The administrative leadership team consists of a district superintendent, principal, two assistant superintendents; one for curriculum and instruction and a second for

business, two assistant principals; one for administrative services and another for student services. Beneath these lines of administration falls the command of six separate directorships and the Dean of Students. Each curriculum department, and in some cases two to three smaller departments, is headed by a department chairperson who reports to the Assistant Superintendent for Curriculum and Instruction.

The targeted high school district employs 269 full-time equivalent faculty, whom are inclusive school personnel, categorized by the district as classroom teachers. The faculty averages 16.3 years of teaching experience. Teachers and administrators with bachelor's degrees as their highest degree earned account for 21.8% of the population, while 78.2% of the teachers and administrators have earned master's degrees and above. Pupil to teacher ratio in the classroom is 14.2:1, which does not include special education teachers.

The facility is housed in one large four-story building with most everything centrally located on campus. The district administrative offices are located at a second site, a west campus which was closed fifteen years ago as a result of declining enrollment. The Driver Education Department is located in the lower level of the main floor, which was part of an addition built for the athletic department. The classroom and simulator phases of the program are taught in adjacent rooms, while the behind-the-wheel phase meets outside the indoor facility near the district vehicle parking lot.

The high school district is a college preparatory public high school with instructional expenditure per pupil reaching $7,075 and operating expenditure per pupil at $13,528 for the 1997-98 school year. The school attained a 98.6% graduation rate during that same time period.

The targeted school's driver education program area consists of four full-time equivalent instructors and one Department Chairman who also doubles as the Technology Education Department Chairman. All four instructors teach in all three

phases of the school's Driver Education program, which includes classroom theory, simulation, and behind-the-wheel (lab).

The context area has become an increasing problem for this school's department program, since the enactment of the Secretary of State's optional Cooperative Driver Testing Program (CDTP), also known as the waiver program. The CDTP involves a road test and evaluation of basic performance skills in the lab phase of the program, given at school rather than the Secretary of State's office, for those who qualify. The assessment tool from the Secretary of State's office has been loosely interpreted throughout the decade the school site has implemented it. What makes this high school the special place that it is and continues to be, is the freedom which teachers are allowed to pursue and use their own style and personality to encourage students to grow and progress. Although this is a great luxury and advantage in teaching the whole child, it can at the same time be problematic. Consistent evaluation is a high priority in this elite stress driven community, making accountability and consistency critical.

The Surrounding Community

The single high school district is located in an area which includes some of the most luxurious homes and wealthiest people in the metropolitan area. The high school district's boundaries completely encompass five surrounding small town villages, and a small corner section of two other towns. Six elementary schools comprise the feeder population into the high school district.

In a recent publication of demographic trends and enrollment projections circulated by the township, it was projected that the township housing would remain selective by income level. The housing was previously shown to be considerably more expensive than other surrounding townships. The 1998 median family income in the township ranged from $110,000 to $236,000, averaging approximately $164,000. During the

same year, median home values ranged from $370,000 to $500,000. These figures were shown as ranges to better represent the average ends of all seven towns.

Enrollment projections and an ongoing demographic study have been the major topics of concern for the township, as well as the most widely surveyed, debated, and divisive issues. The school district at one time consisted of two separate four-year high schools. In the 1960s a new high school was built due to rising enrollment and future space concerns. In 1980, due to declining enrollment, the newest of the two schools was gradually blended back into one four-year high school located at the original campus site. The transition took five years and the closing campus operated as a freshman campus during those transition years. In 1985, the closing of the newer campus was complete, yet the community was less than excited about the new set-up. Total enrollment hovered around 4,000 students, and a building jammed packed with wall-to-wall people.

Throughout the late 1980s and early 1990s, enrollment reached a low of 2,700 students. With enrollment projected to surge by 1995, the district was once again faced with a space issue, estimated to be at epidemic proportions by the end of the decade. In the past few years the community has been involved in referendums to resolve the ongoing space crisis. The closed campus is still owned by the district and the latest proposal recommends reopening the campus and operating it as a freshman-only building with a transition period of roughly two years. This change could potentially affect the Driver Education Department in terms of staffing and transporting students to the selected site of instruction. With the grade level of a student no longer being the number one determinant of having a driving permit, there are and continue to be quite a few freshmen eligible to take the class. This, in the long run, may cause some transportation problems as well as space utilization issues. With the forthcoming building and grade structure changes involving the possibility of teachers commuting,

there will be an increased need for valid and consistent evaluation methods.

National Context of the Problem

The problem of valid assessment tests of performance based tasks has been a concern at the state and national levels (Wiggins, 1998). Consistent, fair, objective measurement on performance tasks using state guidelines with different evaluators calls for an assessment tool that can be reliably used among raters. Data collection designs to calibrate an assessment network and provide opportunities for objective and fair measurements was found to increase inter-rater reliability on performance tasks (Engelhard, 1997). With reliability of scores a major necessity for valid decision making on performance assessments, Moore and Young, (1997) stated:

> Because performance assessments often have relatively few tasks, consist of complex tasks, and employ more subjective judgments of raters, traditional approaches to estimation of reliability fall short. Since performance assessments almost always use one or more raters to assign scores, or categories to those being assessed, one major potential source of unreliability is inter-rater disagreement. (p.3)

To eliminate observer disagreement, observers need to be trained with criterion-related agreement measures used both before and during a study, (Frick & Semmel as cited in Moore & Young, 1997).

Percentage agreement between raters is another possible estimation of inter-rater reliability. One-hundred percent agreement is seen as a high rate of inter-rater agreement, while 0% is seen as low inter-rater reliability (Moore & Young, 1997). It is not uncommon to ask raters to categorize and classify information into a four or five point scale. With a lower number of points on a rating scale, chance agreement increases.

Although many performance assessments do have low inter-rater reliability, several performance assessments showed high levels of rater reliability. When well-defined scoring rubrics with intensive training were used along with ongoing monitoring, high levels of rater reliability were found (Moore & Young, 1997). High rater reliability does not necessarily imply that score reliability is even satisfactory in performance assessments. The measurement literature shows that high rater reliability is possible and achievable with two raters or even just one rater, if scoring guidelines are specific and training adequate for the rater (Moore & Young, 1997). It should be noted that raters may introduce error in examinee scores if inadequately trained toward the rating scale, fatigued, deficient in content area, or when personal beliefs conflict with values adopted by the scoring rubric (Wolfe & Chiu, 1997).

CHAPTER 2

PROBLEM DOCUMENTATION

Problem Evidence

In order to document the extent of multiple rater differences on lab assessments at the targeted site, teacher surveys were distributed at the beginning of the fall semester to all driver education staff who teach driver education full time during the regular school year, as well as those instructors who teach the subject only during the summer session. Interviews with teachers were conducted at department meetings each month regarding assessment concerns. Anecdotal records were collected throughout the semester on lab evaluations from department teachers.

Teacher Survey

Each driver education instructor was given a six question survey consisting of open ended free response questions concerning the current lab assessment procedures used at the targeted site. The questions centered around criteria used for assessment, frequency of assessment, amount of parallelism between program phases and personal satisfaction with current methods of assessment. The data discussed regarding teacher surveys was taken from a sample of six driver education faculty. A majority of teachers felt that lab sessions somewhat paralleled the simulator and classroom phases in respect to time lines during the school year, but were divergent during the summer school session.

All teachers reported that they based their assessments on skill performance criteria with one teacher including knowledge in their response, while another mentioned the state Cooperative Driver Testing Program (CDTP) guidelines.

Teachers reported their feedback to students ranged from, after each session to every second or third driving session, as well as one respondent stating they gave a mid-semester letter grade. All the feedback offered was subjective and recorded in narrative form on students' lab progress notes including both strengths and weaknesses, as well as areas to work on outside of class.

The level of satisfaction with the current assessment practices reported by teachers varied greatly. Although a couple of teachers felt the current practices were adequate for the department, the majority made mention of inconsistency and lack of continuity between instructors in areas of student performance expectations and grade assessment. The survey responses indicated that interpretive differences of criteria and expectations among department members has lent itself to widely apparent inconsistencies in rater reliability.

The final survey question asked for suggestions each might like to see. Responses ranged from none, to none but open to anything, to wanting to implement a method of having all teachers teach the skills the exact same way whether in classroom or lab. Specific criteria for receiving a passing grade on each skill was also reported as a desired change. One teacher responded with a desire to drive with his own classroom students. Currently at the targeted site each student has two different instructors. One for simulation and classroom and a second for behind-the-wheel (lab).

Probable Causes

In order to fully comprehend the probable causes for a lack of a consistent assessment instrument for performance-based behind-the-wheel (lab) tests in driver education, it is important to understand why inconsistencies occur in the assessment

process. The following are causes for inconsistent assessment: multiple rater differences, lack of established criteria, curriculum area, environmental factors, and format.

Multiple Rater Differences

Authentic assessments and performance assessments have certainly increased as alternatives in education and viewed as a better method of measuring what we want students to know. According to Moore and Young (1997), the use of one or more raters in performance assessments, which is the normal standard, can potentially produce inter-rater disagreement with accompanying unreliability. Errors in rater assessment can be present from unfamiliarity, as well as inadequate training using the rater scale (Wolf & Chiu, 1997). At the targeted site, individual performance assessments have generally been performed by one rater only, but without regard to reliability measures.

Lack of Established Criteria

At the targeted site, a portion of the strategic plan and district goals included developing assessment tools within all departments that measure progress toward established goals. Some departments clearly had well established goals and criteria, while others were less formally constructed. According to Wiggins (1998, p. 169), "Current rubrics tend to overvalue specific methods and formats while undervaluing the result." The district goal for departments sought to assess progress toward established goals, not just end results. Without a sense of more established criteria and ongoing assessment, measuring progress towards these goals was inconsistent, if present at all.

Curriculum Area

The targeted high school's curricular area of driver education was at one time partially graded on a pass or fail basis, that being the performance-based lab phase. With only two options, pass or fail, the line could be blurred and anything but clear cut.

The separation in and between students was immeasurable. Popham (1998, pp. 307-308) found that, "Pass or fail grading systems, because they separate students into only two groups, are insufficiently discriminating to contribute all that much to diversity on student grade point averages, hence are not often employed prior to college." Driver Education is classified as a minor subject with accompanying minor credit at the targeted high school. These minor subjects can be viewed with less value and importance in some aspects of the curricular program at the targeted site.

Approximately eight years ago the targeted high school became a participating member school in the Cooperative Driver Testing Program (CDTP). This program allows qualified students to take their driving test for their state license at the high school site with their regular lab teacher. Minimum grade standards must be achieved by the end of the semester to become eligible and qualify to take the CDTP waiver road test. The implementation of the CDTP at the site now mandated performance letter grades in the lab phase of the program, since evidence had to be shown for achievement at the qualifying minimum standard grade. This was the catalyst necessary to be discriminating enough to establish some type of criteria, subjective as they may have been upon initiation. Haydel and Oescher (1995) commented that as the importance of the decision increases from the performance assessment, so should the assessment environment.

Environmental Factors

Performance assessments in the targeted curricular area occur in natural, variable settings. This presents somewhat of a changing environment, with each lab assessment exposed to non-identical factors, some of which occur quite randomly. This alone can be a major cause of inconsistent evaluation, even with one evaluator, but certainly with multiple raters.

Format

The targeted driver education department currently uses an open ended narrative style form for student progress and assessment with criteria subject to interpretation by each individual teacher. Wiggins and McTighe (1998) see backward planning as an alternative to coverage and activity plans. The teachers and evaluators need to know what they want students to be able to do and what evidence will show they have learned it. The show of evidence, as well as criteria for evidence to increase consistency must be consensual among evaluators.

CHAPTER 3

THE SOLUTION STRATEGY

Literature Review

The topics for discussion that have been found to be solutions for the lack of consistent assessment instruments for the performance-based lab (behind-the-wheel) phase of a Driver Education program include: developing appropriate criteria to be included in the assessment, training raters for reliability, assessing students by multi-raters, and planning backwards from outcomes of education to provide a practical framework for designing curriculum. For consistent assessments to take place, appropriate criteria need to be developed.

Developing Appropriate Criteria

Before consistent assessment can take place, teachers must select those performances which provide direct measurement of real performance on important tasks (Haydel et al, 1995). Those skills or performance selected should be able to be measured, as well as provide results which show the measurement of real performance. Should inappropriate criteria be used in an assessment the measurement would not show direct real performance and the reliability would be compromised.

Assessing instruction and performance when including appropriate criteria will lead to better measurement of performance on those tasks. One such tool used for assessing instruction and performance, based on predetermined expectations and

criteria are rubrics. By developing a rubric to include those criteria selected as important performance tasks, teachers can increase their chance of direct measurement of real performance. Rubrics are but one choice of alternative assessment.

By developing appropriate criteria for the assessment, as well as using a performance-based method, the evaluator can gather information on how a student understands and applies this knowledge (Brualdi, 1998). In driver education although instruction is comprised of multiple phases, ultimately the bottom line becomes one's ability to apply that knowledge in a practical setting, that being behind-the-wheel with other road users. The knowledge attainment of theory components of the course is not a clear determinant of successful performance-based application.

According to Arter (1998, p. 6), "Teachers tend to be better at developing rich interesting tasks in which to engage students than they are at developing the criteria that describe quality performance on the task." Both components are needed to make it an assessment. The issue of appropriate criteria surfaces again here. Teachers can create, demonstrate, design and implement wonderfully engaging tasks, rich with substance and creativity, yet be lacking in ability to objectively develop appropriate criteria describing quality performance. If one is unable to describe which criteria would demonstrate quality performance then the assessment would not be consistent or reliable, whether single or multiple raters were used.

While the development of appropriate criteria included in an assessment is necessary for direct measurement of real performance, are teacher and student both clear on criteria for success? Students will surely have difficulty defining successful criteria if teachers are unsure of their own criteria for adequate performance (Saphier & Gower, 1987). Deficiency in this area will result in unreliable results in assessment within each individual evaluator, as well as among multiple raters.

Training Raters for Reliability

Consistent assessments are also contingent on the training of raters for reliability. In order for rater training to be successful, raters should be familiar with the measures they will use, ensure that they understand the sequence of operations they must perform and explain how normative data should be interpreted (Rudner, 1992). When unfamiliarity with measures of operation exist, rater training effects can be minimized. It is imperative that inservice programs or sessions be developed to increase the chances of raters all being familiar, as well as comfortable with various measuring methods used. Interpretation is another area of subjectivity which should be included.

Linn & Burton (as cited in Moore & Young, 1997) found high levels of generalizability after reviewing several performance assessments when well defined scoring rubrics with intensive training and ongoing monitoring during rating sessions was used. This was not the case across-tasks, as generalizability was limited.

The assessment as reliable refers to consistency with which a test measures whatever it is measuring. In other words, reliability equals consistency, according to Popham (1999). Well defined criteria are essential to establishing reliability. If a test is measuring something other than what is intended the test reliability or consistency declines. Gipps (as cited in Reckase, 1997) indicates, "We do not see assessment as a scientific, objective activity, this we now understand to be spurious", and further stated, "Evaluation within the constructivist and naturalistic paradigms rejects the traditional criteria of reliability, validity and generalizability and looks instead for qualities such as trustworthiness and authenticity" (pp. 1-2). Reckase (1997) argues that special procedures for constructing assessment tools containing performance assessment tasks are unnecessary and that current test development methodology can easily be generalized to complex performance assessment tasks without destroying the desirable characteristics of those tasks.

If the proposed use for a performance assessment is learning improvement what, technical qualities should it have to support this goal? As Reckase (1997) states,

> Strictly from an instructional perspective, assessments used for instructional support should provide rich activities that match the goals of instruction and feedback to the student about the accomplishment of the goals. It is important that the student know that someone is paying attention to what they do. For these purposes, perhaps all that is needed is feedback from a credible source. That is, no technical requirements need be met since the assessment is part of the teacher/student interaction and individuals outside that interaction do not need to interpret the results. (p. 6)

If the assessment must be high stakes to motivate the student to perform at their best, Resnick and Resnick (1996) indicate that, "Without incentives for students to engage in the kind of challenging work that complex tasks represent at any grade level, it is unlikely that direct measures for assessment will fully produce the desired effect on learning" (p. 32).

A somewhat less extreme position is presented by Moss (1992) who quotes a personnel communication from Allan Collins: "Collins notes that they (Frederiksen and Collins) have moved away from a sampling model of measurement to a performance model (similar to that used in the Olympic Games), where the quality of the performance and the fairness of the scoring are crucial but where replicability and generalizability of the performance are not" (p. 250). If the performance model of the Olympic Games presented by Moss (as cited in Reckase, 1997) is taken at face value, the goal of instruction becomes a high level of performance on the assessment task in the same sense that the goal for the Olympic athletes is to win their event. All training is focused on improving the likelihood of achieving that goal. Under this model, the technical requirement of the assessment is reliable scoring.

Moss (1992) indicated that within the Olympic scoring format for events such as diving or ice skating, task requirements are well known by all participants in advance, the judges are well trained on very specific rubrics, multiple judges are used, and the high and low judgments may have been dropped to stabilize the averages of the judges' ratings. Performance on the task becomes the goal. Reckase (1997) further states that, "While precision in scoring is clearly needed, whether spread in scores is required depends on whether a mastery model or an individual differences model is used for assessment" (p. 7).

Use of a mastery model implies that the percent of exact agreement is the statistic of choice for evaluating the quality of assessment. However, if detecting differences in level of performance of students is critical, as it is in Moss' (1992) Olympic Games analogy, then it is important that there be sufficient spread of scores to allow relatively fine distinctions in performance to be made. Performance assessment tasks and scoring rubrics should be designed so that score distributions on the tasks include all scoring categories (Reckase, 1997).

Overall assessment ratings may also be a significant factor in the daily assessment or lack thereof in the classroom. According to Stiggins (as cited in Burke, 1994),

> Our current assessment values may also be contributing to inadequate daily assessment of student achievement in some classrooms. Since we have rarely inquired into the quality of teacher-developed tests, offered training in classroom assessment, or included classroom assessment in the Principal's leadership role, we simply do not know how well teachers measure student achievement or how to help them if they need help. (p. xi)

Many researchers have shown that using performance assessments to generalize to other tasks is questionable (Brennan, 1996; Dunbar, Koretz & Hoover, 1991; Shavelson, Baxter & Gao, 1993).

The following options are available for those test developers that desire to produce a performance assessment that yields generalizable results. The first option is to select performance tasks that are at least moderately intercorrelated. Good inter-rater reliability is a necessity in achieving this goal. Pretesting a number of tasks for their intercorrelation may be appropriate for the final selection of those tasks included in the assessment. The potential disadvantage using this method is that the process of task selection might narrow the domain that is being assessed. the second option is to increase the number of assessment tasks administered until the desired level of generalizability is attained.

According to Brennan et al. (as cited in Gao & Colton, 1996) variations in sampling as well as generalizability of performance assessments has resulted in the following indications; "(a) an individual's performance score varies greatly from one task to another, (b) a large number of tasks are needed to obtain a generalizable measure of an individual's performance and (c) well trained raters can provide reliable ratings" (p.59). A high generalizability coefficient means that students will likely be equally capable on other tasks of the same type. High generalizability for a particular performance assessment does not necessarily indicate that the level of performance will generalize to an entire domain. Messick (as cited in Reckase, 1997) states,

> Domain coverage is critical to the construction of performance assessments. If the domain is thought to be fairly unidimensional, defining a continuum of skills, then domain coverage can be demonstrated by showing that the assessment tasks provide information over the range of the continuum that is of interest. (p. 10)

Numerous rater reliability studies have been undertaken to determine whether raters are unduly influencing examinees scores. Wolfe and Chiu (1997) offer two possible frameworks to examine rater effects; normative and criterion-referenced.

In a normative framework, the more common of the two, rater effects are examined

> in the context of the pool of raters from which individual raters are drawn. Hence, a normative framework for examining rater effects describes how much individual raters differ from the "average" rater in the pool. As a result, the normative framework can also be referred to as an agreement framework because we are concerned with how well the ratings of individual raters agree with the ratings assigned by all of the other raters in the pool. (p. 4)

The second framework identified is criterion-referenced in nature. According to Wolf and Chiu (1997), rater effects can be examined in the context of some external point of reference. This reference is assumed to be a valid indicator of the examinee's proficiency. Externally-generated scores are routinely assigned by a benchmark committee although alternative methods may be used, such as taking examinee's scores from another assessment instrument. Criterion-referenced framework depicts rater errors rather than effects by measuring the accuracy of a rater's ratings rather than simply the agreement of those ratings assigned by other raters.

Scores which are based on ratings carry potential threats to their validity. Those being rated may not be performing in their usual manner, thus yielding atypical behavior on the task being assessed. The rater errors may stem from the raters themselves, unintentionally distorting the results. The following five rater effects represent some of those identified and studied: the halo effect, stereotyping, perception differences, leniency/stringency error and scale shrinking (Rudner, 1992).

The halo effect. Raters may form impressions about an individual on one dimension which can carry over to and influence their impressions of that same person on other dimensions. Nisbett and Wilsons' study (as cited in Rudner, 1992) found evidence of the halo effect after making two videotapes of the same professor. In the first videotape the professor acted in a friendly manner towards his class, while in the second he behaved arrogantly. Students who viewed the friendly professor tape rated

the professor more favorably on other traits including physical appearance and mannerisms.

Stereotyping. Impressions that an evaluator forms on an entire group can influence their impressions about an individual group member. In other words a principal might find a driver education teacher to be a safe driver because all driver education teachers are supposed to be safe. The evaluator uses their perceived expectation of what typically has been the norm or general typecast for a group and applies the generalization to all individuals from the group.

Perception differences. An evaluator's current viewpoints as well as past experiences can certainly affect their interpretation of behavior. Dearborn and Simons' study (as cited in Rudner, 1992) found evidence of perception differences when they asked business executives to identify the major problem in a detailed case study. The executives tended to view the problem in terms of their own departmental functions.

Leniency/stringency error. Without adequate knowledge or information to make an objective rating, an evaluator may give scores which are systematically higher or lower as a form of compensation. The rating appears as either extremely lenient or excessively stringent and falls at one end of the scale or the other. The rater error in this case can cause issues with score validity, since either high or low end scores may not be used in the final rating.

Scale shrinking. Raters who refuse to use the end of any scale narrow the given range of scores, thereby increasing the density of scores in the scale's midpoint area and shrinking the scope and size of the scale. Validity most certainly will be compromised in this situation.

In addition to training raters for reliability in performance assessment, rating scales and associated points of degree levels can influence the reliability of the test. Myford et al. (1996) found that rating scales with seven to ten points rather than three

or four showed little appreciable gain in reliability for scales having more than five points. The particular features of the scale were not as important as the knowledge, skills, and motivation of the rater.

Assessment by Multi-raters

The formation and selection process of judges for a review panel consisting of multi-raters encompasses many variables. The choice of rater or judge can have a significant influence on scores. Common issues surrounding the selection of judges include demographics, expert versus interest groups, and split panels.

In identifying many of the issues involving judge selection, Hambleton and Powell (1983) offered the following recommendations to questions accompanying these issues. Demographic variables such as race, sex, age, education, occupation, specialty, and willingness to participate should be considered in the selection process. Credibility may rise with a varied composition of panel members. Whenever possible, review panels should be composed of both experts and representatives from interest groups. The authors argue that review panels should be divided into smaller working groups when the review panel is too large to allow effective discussion to take place. Review panels should also be split if ratings are going to be compared across groups to access reliability or to cross check validity.

The difference between a rater's average and the average of all ratings is called the rater effect. If the rater effect is zero, no systematic bias exists in the scores. Based on the earlier discussed rater errors, the rater effect is rarely zero. According to Rudner (1992),

> If all the judges rate everyone being evaluated, some rater effects may not be a problem: The candidates all realize the same benefit or penalty from the rater's leniency or harshness. The ranks are not biased, and no one receives preferential treatment. However, an issue arises if different sets of multiple raters are used--

> a common situation when scoring essays, accrediting institutions, and evaluating teacher performance. Candidates evaluated by different sets of multiple raters may receive biased scores because they drew relatively lenient or relatively harsh judges. (p. 2)

Individuals administering and scoring tests can cause errors with carelessness or unfamiliarity. Gall, Gall, and Borg (1999) found that the presence of these errors can be determined by having several people administer the same test to the same sample. The degree of reliability, calculated as a reliability coefficient, is the inter-rater reliability or inter-observer reliability. When analyzing tests with variation in severity of rater, Longford (1996) found that taking between rater differences into account was of high importance.

Backward Planning

If the desired outcome of assessment is improved performance, the techniques of measurement must be accompanied by quality feedback to the learners. Wiggins (1998, p. 43) states, "The feedback needs to be of two kinds: in addition to better feedback after the performance, feedback must also be provided during (concurrent with) the assessment activities." According to Johnson (1996), regarding performance assessments, teachers must plan backwards from outcomes of education to shift the paradigm of curriculum-instruction-testing to a new, fluid design.

In a study by Wiggins & McTighe (1998), using a multi-faceted approach, with six facets of understanding combined with backward decision provided a practical framework for designing curriculum, assessment, and instruction. The efficiency and effectiveness of assessment is influenced by many variables. While increased student load and financial constraints make efficient assessment even more difficult to maintain, it becomes imperative for assessments to be fair, valid, reliable, and effective for instruction and performance improvement (Smith, Brown, and Race, 1996).

The topics discussed for solutions to a lack of consistent assessment instrument for the lab phase of driver education; developing appropriate criteria to be included, training raters for reliability, assessing students by multi-raters, and planning backwards from outcomes of education to provide a practical framework for designing curriculum will be addressed in the intervention. Results of the project will be discussed in Chapter 4.

Project Objective

As a result of developing a performance-based scoring assessment during the period of September, 1999 through January, 2000, the targeted high school will implement and adopt this assessment instrument as a Driver Education Department standard. This will be monitored by teacher interviews, teacher surveys, and anecdotal records.

Process Statements

In order to accomplish the project objective, the following processes are necessary:

1. Evaluate the current assessment system
2. Establish appropriate and desired criteria to use for assessment
3. Develop an instrument aligned with curriculum for assessment
4. Pilot the instrument
5. Gather feedback from pilot attempt
6. Revise instrument
7. Collect data on reliability and validity

Project Action Plan

I. Collect problem evidence data (End of August)
   - A. Conduct interviews with the department members
   - B. Keep anecdotal records
   - C. Application of past assessment tools for current system's goals and criteria

II. Gather departmental feedback regarding desired criteria (September)

A. Gather input from department meetings (Early September)

B. Develop instrument for criteria assessment (September)

C. Get feedback from department members

III. Pilot instrument (month of October)

A. All lab teachers try out instrument for month of October

B. Data collected from new instrument

C. Interview other teachers regarding concerns and/or problems, give staff a Plus-Minus-Interesting (P.M.I.) feedback sheet to complete

IV. Revise instrument (Last week of October which concludes 1st academic quarter)

A. Gather data from teachers and fine tune

V. Collect data on two targeted Driver Education classes for reliability and validity using final revised form (November through January-second academic quarter)

A. Data collected from two lab classes by multiple raters

1. Both raters observing same lab sessions using same student sample

Methods of Assessment

In order to assess the effects of the performance-based scoring instrument in Driver Education lab classes, teacher interviews, teacher surveys, and anecdotal records will be gathered and reviewed on a periodic basis.

CHAPTER 4

PROJECT RESULTS

Historical Description of the Intervention

The objective of this project was to develop a performance-based scoring assessment for the behind-the-wheel (lab) phase of the Driver Education program which the targeted high school would implement and adopt as the department standard. Teacher surveys, anecdotal records, as well as design and implementation of a pilot assessment instrument were selected to effect the desired changes.

Teacher Interviews and Surveys

Teacher interviews involving past and present members of the Driver Education department were initiated in September 1999, to determine if the department members felt that a scoring assessment was necessary. A written teacher survey was then compiled and distributed later that month, a sample of which can be found in Appendix A. Using department meetings as a forum, information and criteria was gathered to develop a pilot instrument for criteria assessment.

Scoring Instrument

After piloting the scoring instrument for one month and reinterviewing the teachers, in addition to distributing a Plus-Minus-Interesting (P.M.I.) survey to address further concerns of the process, data was collected and the assessment instrument was revised. A sample copy of the assessment instrument can be found in Appendix B. Original plans called for the intervention to follow during the second quarter of the first

semester during the 1999-2000 school year. The intervention was rescheduled for the summer session of 2000 due to a change in the researcher's class schedule. The intervention occurred during a four week time span equivalent to one academic quarter. With the change in intervention schedule, adjustments in the number of raters used and student sample were necessary to complete the action plan. Two class periods of students were observed; eight students in one class and seven students in the second class. The classes were divided into four lab groups, each comprising two students with the exception of one group which was composed of a single student. The assessment sessions for each lab group from both classes occurred every fourth day of the month-long summer session, with two raters independently observing and assessing the same lab session using the revised assessment instrument.

Presentation and Analysis of Results

In order to assess the effects of developing and implementing a performance-based assessment for the behind-the-wheel (lab) phase of Driver Education with multiple rater consistency, assessments were conducted and scores recorded throughout the intervention. These data were aggregated by each rater's individual score on four distinct skill performance categories, assessed during three separate driving sessions for each of the fifteen students and are presented in Table 1.

During data analysis, the researcher found patterns of consistency had emerged between raters. Of the total 180 rated events, 123, or 68.3% resulted in identical scores between both raters using the four point rating scale. A one point scoring difference occurred in 54 of the 180 rated events, accounting for 30.0% of the total events assessed. A minimum two point difference between raters was evident on three occurrences yielding the final 1.7% of rated events. The 54 events showing a one point difference between raters revealed rater number 1 (R1) to be the high scorer 59.3% of the time, while rater number 2 (R2) was high scorer 40.7% of the time.

Table 1

Multiple Rater Assessment of Student Driver Education Lab Performance

	Session #1								Session #2								Session #3							
	Turns		Traffic Controls		Lane Changes		Right of Way		Turns		Traffic Controls		Lane Changes		Right of Way		Turns		Traffic Controls		Lane Changes		Right of Way	
	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2
Student 1	3	4	4	4	4	4	4	4	3	4	3	4	4	4	4	4	4	4	4	4	4	4	4	4
Student 2	3	2	2	3	3	3	3	3	3	3	4	3	3	3	4	4	4	3	3	4	4	4	4	4
Student 3	4	4	4	4	4	4	4	4	3	4	4	4	3	3	4	4	4	3	4	4	4	3	3	3
Student 4	3	3	4	4	3	3	4	3	3	3	3	3	3	3	3	3	4	4	3	4	4	4	4	4
Student 5	3	2	4	4	3	3	4	4	3	3	3	3	2	3	2	2	3	3	3	3	3	4	3	4
Student 6	3	3	4	4	3	3	3	3	4	3	4	3	3	2	4	3	4	4	4	4	3	4	4	4
Student 7	4	3	3	4	3	3	3	4	4	3	4	4	3	3	4	4	3	3	3	4	4	4	4	4
Student 8	4	3	3	4	4	3	4	2	3	2	2	2	3	3	2	2	4	3	3	3	3	4	4	3
Student 9	4	4	4	4	4	3	4	4	4	4	4	2	4	4	4	3	4	4	4	4	4	4	4	4
Student 10	4	4	4	4	4	4	4	4	4	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Student 11	3	4	3	4	3	3	3	3	4	3	4	3	3	3	3	2	4	3	3	3	4	3	3	3
Student 12	4	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	3	4	4	4
Student 13	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Student 14	4	3	4	4	3	3	4	4	4	3	4	2	3	3	4	4	3	2	4	4	3	3	3	3
Student 15	3	4	4	4	4	4	4	4	4	4	3	4	4	3	4	4	4	3	4	4	3	4	4	4

Note. R1 = this researcher
R2 = an independent rater
4=Good
3=Average
2=Fair
1=Weak/Poor

BEST COPY AVAILABLE

Rater number 1 (R1) assessed the high score 100% of the time when raters showed at least a two point difference.

Although the action plan called for the assessments to be completed independent of other raters, on occasion following a driving session, one rater attempted to get feedback from the other rater regarding their event rating. There were times the researcher observed some uncertainty or vagueness on the part of one or both of the raters.

The results and analysis of the action plan to develop, pilot, and implement an assessment instrument for lab performance in Driver Education has been discussed. The conclusions drawn and recommendations for improving the implementations will be discussed in the following section.

Conclusions and Recommendations

Based on the presentation and analysis of consistency between multiple raters on lab performance assessments in Driver Education the targeted site now has an assessment instrument which showed high levels of consistent scoring between raters. The intervention which included developing appropriate criteria to be included in the assessment, training raters for reliability, assessing students by multiple raters, and planning backwards from outcomes of education for designing curriculum all contributed to the success of developing, piloting, and implementing a department assessment instrument.

Teacher Surveys

The teacher surveys in the form of a P.M.I. worksheet, which may be found in Appendix C, showed more pluses than minuses following use of the assessment rubric. It was a concise method for each teacher to individualize what they felt positive, negative or indifferent about concerning the assessment tool. Collegiality appeared to improve as the teachers were working together towards a common goal.

Developing Appropriate Criteria

Anecdotal notes from department meetings regarding curriculum goals on lab assessments, including district and state goals, provided a framework for selecting appropriate criteria to include in the targeted site's assessment. The high school's participation in the CDTP waiver program further narrowed the focus for appropriate criteria. The researcher used the state CDTP guidelines to show measurement results of real performance.

The goal of this research project was to improve rater reliability on performance based Driver Education lab assessments through the consistent use of an assessment instrument. Developing appropriate criteria for assessment, piloting the instrument, revising the instrument, and implementing its use during selected classes were all employed to achieve this goal. This particular plan appeared to have a positive impact on the targeted site at the local level. Based on the fact that valid assessment tests of performance based tasks has been a concern at the state and national levels, a consistent measurement on performance tasks with multiple raters is imperative for assessments to be considered reliable. Those performances selected must provide direct measurement of real performance on important tasks or reliability is compromised.

The researcher endorses this intervention with some modifications. The time frame for rater training and piloting the assessment needs to be extended to allow for sudden changes in the class schedule, teacher's schedules, and school calendar. The additional time would allow the researcher to work around those unplanned deviations from the regular schedule with maximum flexibility and a minimum of further disruptions. The summer session proved to be an easier schedule to adapt to for performing lab assessments with fewer changes or adjustments to deal with. Decreased enrollment, lighter schedules, and a compacted day during the summer

session had a positive effect on implementing the intervention. This type of intervention is extremely time consuming and requires cooperation and communication among participants. Consider personalities before committing to a long range project of this nature. Since contact with colleagues is frequent and often intense, the researcher recommends establishing a good working environment up-front. The success of a project such as this necessitates that the the entire action research team be focused on carrying out the action plan as outlined, independent of their own biases.

This researcher recognized a benefit to the multiple rater approach because students could be assessed by raters similarly trained in the consistent use of the scoring instrument. Students and parents both benefit by receiving more detailed feedback through periodic progress reports assessing their present lab performance. In addition, students and parents will have specific performance criteria to guide them through their driving sessions at home in order to develop and practice those skills necessary to safely and successfully pass their driving tests. With consistent results between raters using the scoring rubric, a student can be assessed by any trained rater with comparable results, lessening the chance of introducing rater error. Accountability and consistency are educational issues confronting our nation in the 21st century, much of which falls into the lap of the teacher standing at the front line. Improvement, growth and continual progress, which can be measured and documented, are essential to positive student outcomes.

REFERENCES

Arter, J. (1998, April). Teaching about performance assessment. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Brennan, R. L. (1998). Misconceptions at the intersection of measurement theory and practice. Educational Measurement: Issues and Practice, 17 (1), pp. 5-9, 30.

Brualdi, A. (1998). Implementing performance assessment in the classroom. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation. (ERIC Document Reproduction Service No. ED 423 312)

Burke, K. (1994). The mindful school: How to assess authentic learning (Rev. ed.). Arlington Heights, IL: IRI/Skylight Training and Publishing.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. Applied Measurement in Education, 4, 289-303.

Engelhard, G., Jr. (1997). Constructing rater and task banks for performance assessment. Journal of Outcome Measurement, 1, 19-33.

Gall, J. P., Gall, M. D., & Borg, W. R. (1999). Applying educational research: A practical guide (4th ed.). New York: Addison Wesley Longman.

Gao, X., & Colton, D. A. (1996, April). Evaluating measurement precision of performance assessment with multiple forms, raters, and tasks. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Hambleton, R. K., & Powell, S. (1983). A framework for viewing the process of standard setting. Evaluation and the Health Professions, 6 (1), 3-24.

Haydel, J. B., Oescher, J., & Banbury, M. (1995, April). Assessing classroom teacher's performance assessments. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Johnson, B. (1996). The performance assessment handbook: Vol. 2. Performances and exhibitions. Designs from the field and guidelines for the territory ahead [Abstract]. Larchmont, NY: Eye on Education. (ERIC Document Reproduction Service No. ED 421 539)

Longford, N. T. (1996, Fall). Reconciling experts' differences in setting cut scores for pass-fail decisions. Journal of Educational and Behavioral Statistics, 21 (3), 203-213.

Moore, A. D., & Young, S. (1997, October). Clarifying the blurred image: Estimating the inter-rater reliability of performance assessments. Paper presented at the annual meeting of the Northern Rocky Mountain Educational Research Association, Jackson, WY.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. Review of Educational Research, 62 (3), 229-258.

Myford, C. M. (1996, April). Constructing scoring rubrics: Using "facets" to study design features of descriptive rating scales. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

New Trier Township High School District 203. (1999, November). Illinois School Report Card, 1999 (9-12 D version). Winnetka, IL.

Popham, W. J. (1999). Classroom assessment: What teachers need to know (2nd ed.). Boston: Allyn and Bacon.

Reckase, M. D. (1997, March). Statistical test specifications for performance assessments: Is this an oxymoron? Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Resnick, D. P., & Resnick, L. B. (1996). Performance assessment and the multiple functions of educational measurement. In M. B. Kane & R. Mitchell (Eds.), Implementing performance assessment: Promises, problems, and challenges. Mahwah, NJ: Lawrence Erlbaum Associates.

Rudner, L. M. (1992, December). Reducing errors due to the use of judges. American Institutes for Research. Washington, DC.

Saphier, J., & Gower, R. (1987). The skillful teacher: Building your teaching skills. Carlisle, MA: Research for Better Teaching.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variation of performance assessments. Journal of Educational Measurement, 30 (3), pp. 5-8, 15.

Smith, B., Brown, S., & Race, P. (1996). 500 tips on assessment. London: Kogan Page Limited.

Wiggins, G. (1998). Educative assessment: Designing assessments to inform and improve student performance. San Francisco: Jossey-Bass.

Wiggins, G., & McTighe, J. (1999, February). A designer's handbook: Understanding by design, tests. Paper presented at the New Trier Township High School Winter Institute [Handbook]. Winnetka, IL.

Wolfe, E. W., & Chiu, C. W. T. (1997, March). Detecting rater effects with a multi-faceted rating scale model. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

APPENDICES

APPENDIX A

DRIVER EDUCATION STAFF SURVEY

1. Do your lab (behind-the-wheel) lessons parallel the class and simulator content areas and time-lines?

2. What criteria do you base your assessments on?

3. What feedback do you offer the student regarding their daily performance in the lab phase?

4. How do you assess your lab students on a regular basis? How frequently?

5. How satisfied are you with the current assessment used for lab students in our Department?

6. What changes would you like to see implemented?

APPENDIX B

REVISED ASSESSMENT INSTRUMENT

	4	3	2	1
	GOOD	AVERAGE	FAIR	WEAK/POOR
Turns (Right - Left)				
Traffic Controls (Signs - Signals)				
Lane Changes - Blind Spot Checks				
Right of Way				
Stops/Braking				
Speed Control				
Steering				
Backing/Turnabouts				
Space Cushion - Following Distance				
Attention - Interactions With Other Roadway Users				
Attitude/Cooperation				
Parking (Angle/Perpendicular/Hill/Parallel)				

APPENDIX C

(P. M. I.) SURVEY SHEET

++

PLUS +

MINUS -

INTERESTING *i*

*i i i

APPENDIX D

SAMPLE LETTER TO PARENTS

Dear Parents and Students,

As a requirement for completion of my Masters Degree in Teaching and Leadership from St. Xavier University, I will be conducting an Action Research Project in Driver Education classes during the 2000 summer school session. The purpose of this study is to examine the current assessment tools used to evaluate how well your son or daughter applies the driving skills during the behind-the-wheel (lab) phase of the program. New assessment measures will be developed and implemented in order to adopt a uniform assessment system throughout the department. Assessment is a part of the current Driver Education curriculum. This project aims to improve current practices. Documentation will include lab write-ups of student performance and interviews with other Driver Education teachers.

Benefits for students and parents include more detailed feedback for students through periodic progress reports. In addition, students and parents will have detailed criteria in order to know what skills to practice and develop in order to successfully pass their driving tests.

Strict confidentiality will be maintained while all data and results are collected and reported in my Final Project completed by May 2001. Involvement in the study is on a voluntary basis, and in no case will a student be penalized for declining to participate.

In order to include your child's results in my report, I need your consent. Please indicate your preference for your child's participation in the data collection for the Action Research Project, along with your signatures on the following page and return it to your Driver Education teacher.

Thank you for your consideration, time and involvement in the upcoming project. It is sincerely appreciated. If you have any questions or concerns please do not hesitate in contacting me.

Sincerely,

Driver Education Teacher

YES / NO - My child, _____ can be included
(NAME) - PRINT

in the data collection for the Action Research Project.

X_____ _____

(PARENT SIGNATURE) **(DATE)**

X_____ _____

(STUDENT SIGNATURE) **(DATE)**